50 Cd

A34272 PCT USA - 072854.0120

PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant	:	Johnson et al.
Serial No.	:	09/869,091
Filed	:	June 20 2001
For	:	DATA SWITCHING METHOD AND APPARATUS

CLAIM FOR PRIORITY UNDER 35 U.S.C. § 119

I hereby certify that this paper is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Assistant Commissioner for Patents, Washington, D.C. 20231, on:

August 6, 2001
Date of Deposit

Ronald B. Hildreth
Attorney Name

19,498
PTO Reg. No.

Signature

August 6, 2001
Date of Signature

Assistant Commissioner for Patents
Washington, D.C. 20231

Sir:

A claim for priority is hereby made under the provisions of 35 U.S.C. § 119 for the above-identified PCT application based upon Great Britain application 9828144.7 filed December 22, 1998, and International Application PCT/GB99/03748 filed November 10, 1999.

Respectfully submitted,

Ronald B. Hildreth
Patent Office Reg. No. 19,498

(212) 408-2544
Attorney for Applicants

Baker Botts L.L.P.
30 Rockefeller Plaza
New York NY 10112

CHARTER MARK

The Patent Office

4

PCT/GB99/03748

09/869091

INVESTOR IN PEOPLE

The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ

PRIORITY DOCUMENT

SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

GB99/3748

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.

Signed

Dated 22 NOV 1999

ents Form 1/77

ents Act 1977
(Rule 16)

The Patent Office

Request for grant of a patent

(See the notes on the back of this form. You can also get an explanatory leaflet from the Patent Office to help you fill in this form)

The Patent Office

Cardiff Road
Newport
Gwent NP9 1RH

1. Your reference — M98/0660/GB

2. Patent application number *(The Patent Office will fill in this part)* — 22 DEC 1998 — **9828144.7**

3. Full name, address and postcode of the or of each applicant *(underline all surnames)*

   Power X Limited

   Stafford Court
   145 Washway Road
   Sale
   Cheshire M33 7PE

   Patents ADP number *(if you know it)* — 6803253002

   If the applicant is a corporate body, give the country/state of its incorporation — Great Britain

4. Title of the invention — Data Switching Apparatus

5. Name of your agent *(if you have one)* — McNeight & Lawrence

   "Address for service" in the United Kingdom to which all correspondence should be sent *(including the postcode)*

   Regent House
   Heaton Lane
   Stockport
   Cheshire SK4 1BS

   Patents ADP number *(if you know it)* — 0001115001

6. If you are declaring priority from one or more earlier patent applications, give the country and the date of filing of the or of each of these earlier applications and *(if you know it)* the or each application number

   Country | Priority application number *(if you know it)* | Date of filing *(day / month / year)*

7. If this application is divided or otherwise derived from an earlier UK application, give the number and the filing date of the earlier application

   Number of earlier application | Date of filing *(day / month / year)*

8. Is a statement of inventorship and of right to grant of a patent required in support of this request? *(Answer 'Yes' if:*
   a) any applicant named in part 3 is not an inventor, or
   b) there is an inventor who is not named as an applicant, or
   c) any named applicant is a corporate body.
   See note (d))

   Yes

9. Enter the number of sheets for any of the following items you are filing with this form. Do not count copies of the same document

Continuation sheets of this form

Description 8

Claim*(s)* -

Abstract -

Drawing*(s)* 2 x2

10. If you are also filing any of the following, state how many against each item.

Priority documents

Translations of priority documents

Statement of inventorship and right to grant of a patent *(Patents Form 7/77)*

Request for preliminary examination and search *(Patents Form 9/77)*

Request for substantive examination *(Patents Form 10/77)*

Any other documents *(please specify)*

11. I/We request the grant of a patent on the basis of this application.

Signature McNeight & Lawrence

Date 21/12/1998

12. Name and daytime telephone number of person to contact in the United Kingdom

McNEIGHT & LAWRENCE
0161 480 6394

Warning

After an application for a patent has been filed, the Comptroller of the Patent Office will consider whether publication or communication of the invention should be prohibited or restricted under Section 22 of the Patents Act 1977. You will be informed if it is necessary to prohibit or restrict your invention in this way. Furthermore, if you live in the United Kingdom, Section 23 of the Patents Act 1977 stops you from applying for a patent abroad without first getting written permission from the Patent Office unless an application has been filed at least 6 weeks beforehand in the United Kingdom for a patent for the same invention and either no direction prohibiting publication or communication has been given, or any such direction has been revoked.

Notes

a) If you need help to fill in this form or you have any questions, please contact the Patent Office on 0645 500505.

b) Write your answers in capital letters using black ink or you may type them.

c) If there is not enough space for all the relevant details on any part of this form, please continue on a separate sheet of paper and write "see continuation sheet" in the relevant part(s). Any continuation sheet should be attached to this form.

d) If you have answered 'Yes' Patents Form 7/77 will need to be filed.

e) Once you have filled in the form you must remember to sign and date it.

f) For details of the fee and ways to pay please contact the Patent Office.

DATA SWITCHING APPARATUS

This invention relates to a digital switch which takes application data from numerous input sources and routes it to numerous destination outputs.

Fig 1. shows a generalisation of such a concept. Data arriving on input ports 1..n is routed via the switch to output ports 1..n. For an input N to transfer data to an output **M** the switch establishes a 'connection' between N and M. The connection generally remains for the duration of the data transfer at which point it may be broken and the output allowed to be connected to another input. Data is typically transferred in 'cells'.

Because there are numerous inputs competing for numerous output ports the possibility of contention occurs. The output port can be considered to be a resource that must be shared amongst multiple inputs. This means that a particular input may not be able to connect to a particular output because that output is already in use i.e. is already connected to another port. It is also possible that more than one input may be requesting a connection to the same output. In either case the result is the need for the cells to be queued **(buffered)** until the relevant resource becomes available.

Cells can be stored in several areas in the switch; the input, the output and centrally. Most switches use a combination of all three. It is generally considered that output buffering provides the most efficient way for handling traffic shaping i.e. the profile of the release of cells from the switch. However, output buffering places severe requirements on the actual storage device used to create the buffer. This is because the buffer is shared amongst multiple outputs which means that the storage devices must be very high performance. However, at very high data rates current technology limits the use of output buffers.

With intelligent distributed scheduling mechanisms it is possible to create an input buffered switch which meets the same traffic shaping efficiency of its output buffered counterpart. The use of input buffers is preferred for several reasons. Input buffering requires smaller buffers, which can have relatively low performance and therefore be cheaper.

When cells are queued at the input there is the possibility of contention arising through the phenomena of Head Of Line (*HOL*) blocking. This generally occurs when First In First Out (*FIFO*) queue mechanisms are used. With FIFO queues the cell at the head of the queue is the only one that can be chosen for delivery through the switch. Now, consider the case where an input port has three cells c1, c2, c3 stored such that c1 is at the head of the queue with c2 stored next and c3 last with cell c1 destined for port N and cell c2 destined for port N+1. Now port N is already connected to port N-1 therefore c1 cannot be switched, however port N+1 is unconnected and therefore c2 could actually be delivered. However, c2 cannot get out of the FIFO because it is blocked by the HOL.

An intelligent approach to the solution of HOL blocking is the use of Virtual Output Queues (VOQ). Using VOQs the cells are separated out at the input into queues which map directly to their required output destination. They can therefore be effectively described as being output queues, which are held at the input i.e. ***Virtual*** Output Queues. Since the cells are now separated out in terms of their output destination they can no longer be blocked by the HOL phenomena.

According to the invention there is provided a method of handling packets of information through a data switch comprising input traffic controllers, ingress routers, a memoryless cyclic switch fabric, egress routers and output traffic controllers all under the control of a switch master controller and interconnected such that each input line connected to the data switch is terminated on a traffic controller arranged to convert the input line protocol information packets into fixed length cells having a header defining the data switch destination router and output traffic controller together with message priority information arranged such that each ingress router serves a group of traffic controllers *characterised in that* the ingress router includes a set of input buffers one for each input line and a set of virtual output queue buffers, one for each output traffic controller from the data switch, and in which the method comprises on the arrival of a cell from a traffic controller the ingress router examines the cell header and places it in the appropriate virtual output queue and generates a request for transfer message consisting of the destination traffic controller address and a message priority code which is passed to the data switch master controller, the master controller schedules the passage of the cells across the switch fabric by interconnecting a specific ingress router to a specific egress router for each switch fabric cycle in accordance with a first arbitration process the ingress router selecting from the appropriate virtual output queue the cell at the head of the queue for passage across the data switch to the appropriate output traffic controller in accordance with a second arbitration process.

According to the invention there is provided a data switch for handling packets of information comprising input traffic controllers, ingress routers, a memoryless cyclic switch fabric, egress routers and output traffic controllers all under the control of a switch master controller and interconnected such that each input line connected to the data switch is terminated on a traffic controller arranged to convert the input line protocol information packets into fixed length cells having a header defining the data switch destination router and output traffic controller together with message priority information arranged such that each ingress router serves a group of traffic controllers *characterised in that* the ingress router includes a set of input buffers one for each input line and a set of virtual output queue buffers, one for each output traffic controller connected to the data switch, and in which on the arrival of a cell from a traffic controller the ingress router examines the cell header and places it in the appropriate virtual output queue and generates a request for transfer message consisting of the destination traffic controller address and a message priority code which is passed to the data switch master controller, the master controller schedules the passage of the cells across the switch fabric by interconnecting a specific ingress router to a specific egress router for each switch fabric cycle in accordance with a first arbitration process and the ingress router selects from the appropriate virtual output queue the cell at the head of the queue for passage across the data switch to the appropriate output traffic controller in accordance with a second arbitration process.

There is also the question of Quality Of Service (QoS) to address. Different input sources have different requirements in terms of how their data should be delivered. For example voice data must be guaranteed a very tightly controlled delivery service whereas computer data can be more relaxed. To accommodate these requirements the concept of priority can be used. Data is given a level of priority, which changes the way the switch

deals with it. For example consider two cells in different VOQs c1 and c2 which are both requesting to go to the same output. Although either could be selected only one can be delivered. The cell with the 'highest' priority is chosen. This decision making process is referred to as **Arbitration**. It is not only priority which can be a factor in the arbitration process. Another example would involve monitoring the length of the VOQs and also using them as a determining factor. It should also be noted that as switches become faster and larger then a more intelligent approach to arbitration needs to be sought. The ideal solution is for a **distributed arbitration mechanism** where there exists levels of arbitration right through the switch from the core right out to the inputs. Using such a mechanism arbitration can be very finely tuned to cater for the most demanding QoS requirements.
By using buffers switches run the risk of losing cells i.e. the buffer overflows. To overcome this problem and also to efficiently size the buffers the concept of backpressure flow control across the switch can be employed. Using backpressure an output can inform the input that is connected to it that it is filling too quickly and is about to lose cells. The input can now back off or slow down the rate at which it is sending the cells and therefore reduce or completely eliminate the risk of cell loss.

This patent describes the implementation of a high-speed digital switch for use in any area in which high speed high performance digital communications is required. Typically this definition covers at least the Data Communications sector and the Cluster Computing sector.

The architecture can be abstracted as shown in Fig 1. Digital data arrives over *Line Ends* ① on any one of the input ports 1-n and is switched through the *switch core* to any one of the output ports 1-n where it leaves the switch on another Line End. Data is carried on the line ends via any one of several *Line* protocols. Typical examples of these Line protocols are Synchronous Optical Network (*SONET*), Asynchronous Transfer Mode (*ATM*), Frame Relay and Ethernet. These line ends typically connect to either a LAN (*Local Area Network*) or WAN (*Wide Area Network*) network environment.

The *switch core* shown in Figure Two. is a cell based switch that operates on data from the traffic controllers which has been packetised into streams of equal 'cell' size.
A 'cell' is a fixed size unit, which forms part of a larger communication unit, which lives at the traffic controller level and is referred to as a 'frame'.
The main components of Fig 2 are: -

CSIX	Common Switch Interface.
Line ends	Ingress and egress ports onto a LAN/WAN network environment.
Terachannel Core	The Terachannel switch core (TERACORE).
Terachannel Core Port 1.. n (tcp 1..n)	Ingress and egress ports for the access to the Terachannel core.
Traffic Controllers (TC)	

On receiving data over the line end the TC stores this data in a buffer which is sometimes referred to as a congestion buffer. The data is sent and received on the line end in protocol dependant units. For example in ATM the data is sent in units of 53-byte ATM cells. Whereas the Ethernet protocol sends data as frames which can be up to several hundreds of bytes long. Line end protocols send these data units with control information in the form of a header, which indicates how the data is to be switched. This header information includes such things as the destination address and priority.
The TC has a mapping of all the other TCs that are connected to the switch, which is usually held in the form of an address translation lookup table. This lookup table is used to calculate which Terachannel Core Port (*tcp*) the data needs to be switched to.
The TC sends and receives data from the Teracore in the form of CSIX packets across the CSIX interface. A CSIX packet contains user payload and control information in the form of a CSIX header. The Teracore uses this header information to determine how the CSIX packet should be switched.

The Teracore comprises three main components: - Router, Master and Crossbar Matrix.

A diagram of the Teracore can be found in Figure 3.

Access into and out of the Teracore is achieved via the Router. When a TC wishes to send data to another TC it does so by sending a CSIX packet across the CSIX interface to the Router and in so doing requests a connection through the core to the destination TC. The Router uses the CSIX header information to determine how the CSIX packet should be switched and establishes a connection in the Teracore between the two TCs to allow the transfer of the CSIX packet to occur.
The CSIX header contains several bits of control information including: -

a. Destination TC address.
b. Priority of the requested transfer.

There are multiple Routers and there can be multiple TCs per Router.
There is only one Master control function, which may take the form of a single device or may be distributed over several Master devices.
The Master is responsible for establishing connections between Routers in the switch core, which in this case is the Crossbar matrix (*CM*). Connections can be one of two types either **transient** or **pseudostatic**. A transient connection exists for the duration of a switch cycle. At the end of each switch cycle the switch is reconfigured by the Master to establish another set of connections. A pseudostatic connection exists over multiple switch cycles and is under the control of a higher level program e.g. a switch network manager.
Routers wishing to connect to each other make requests for connection to the Master across the Master Router Interface (MRI).
Since there are multiple Routers requesting simultaneous connections there is a probability that some of the Routers will be requesting the same connection. This being the case the Master must arbitrate over the requesting Routers to establish which Routers

will be giving connections. This Master arbitration function represents one aspect of the **distributed arbitration** mechanism, which the Terachannel incorporates. Another aspect of the distributed arbitration mechanism lives in the Routers. Once a Router has been granted a connection it must establish which cell to transfer, this is also decided via arbitration.

The Master carries out its arbitration process using several levels of information. The Master uses a combination of information contained in the Bandwidth Allocation Table (*BAT*) and the Priority Allocation Table (*PAT*) to carry out the arbitration process. The BAT holds a set of weights, which are allocated to each possible connection through the CM. One possible technique for achieving this function is called the 'Probabilistic Masking For Bandwidth Allocation' which is represented in the diagram by the MASK block, the details of which are covered in another patent.

When the Master can establish the connection it notifies all the ingress Routers via the MRI in the form of a 'grant' notification it also notifies Routers which are about to receive a cell.

On receiving the grant notification the ingress Router takes the required cell out of the VOQ and serialises it through the Parallel to Serial converter (PS) for transfer across the CM.

The egress Router has been informed that it should expect a cell via the MRI. The cell is delivered into the Serial to Parallel (SP) converter from where it is delivered to output queues oq*n by the Egress Control Unit (*ECU*).

As can be seen in Fig 3, a set of N low bandwidth ports ①, which are connected to the TCs, fill one of N input queues **iq*n**. An Ingress Control Unit (*ICU*) extracts the destination TC addresses from the cells in the input queues and transfers them into a set of M virtual output queues (*VOQ*). There is one virtual output queue for each low-bandwidth output port in the switch.

The ingress multiplexer contains an NxM entry Ingress Port Table (IPT) which defines how its bandwidth to a particular egress port (via a particular virtual output queue) is distributed across the input ports. This table is used by the ICU to determine when (and to what degree) to exert backpressure to the data source resolved down to an individual virtual output queue.

The ingress multiplexer sends control information over the MRI to the Master indicating the state of the virtual output queues (*connection requests*). The Master responds to the Routers over the MRI with a sequence of connections, which it will establish between the Routers (*connection grants*). The ingress multiplexer must now allocate the bandwidth to each egress demultiplexer provided by the Master across the VOQs associated with each egress demultiplexer. The ingress multiplexer contains an Interconnect Link Control Unit (*ILCU*) which implements this function by scheduling cells from the virtual output queues across the high bandwidth link to the CM according to an M entry Egress Port Table (EPT).

Data Transmission through the Terachannel

Data is handled throughout TeraChannel in fixed length cells. The reason for this is that it is much easier for the switch to operate if all the inputs are switched simultaneously, and this is only possible if fixed length cells are used. In practice, there are slight variations in the format of the cells, due to the need to include steering information in headers at various points.

Figure Four shows the flow of data through the switch fabric.

Packets received from a line end are, where necessary, segmented in a traffic manager and formed into fixed-length cells of the correct format to be transferred over CSIX. At the ingress StarRouter, arriving cells are examined and placed in the appropriate queue. To match the virtual output queue requirements, there are up to 512 unicast queues (for unicast cells destined for up to 128 traffic managers, each with four channels), up to 64 multicast queues (for 16 ports, each with four channels) and one broadcast queue. In the diagram, the cell has been placed in one of the unicast queues.

The arrival of a cell triggers a 'request to transfer' to the StarMaster. The cell will be held in the queue, moving up to the head of its queue, until this request is granted. When the request is granted, the StarMaster informs the egress StarRouter to expect a cell. The cell is transferred through the memoryless switch fabric and into a buffer in the egress StarRouter. There is one egress buffer per Traffic Manager and arriving cells are examined and placed in the appropriate queue.

This egress buffering caters for the gear change between the speed of operation of the switch fabric and the speed of the CSIX. The cell is transferred to the traffic manager and, where necessary, re-assembled into a packet before onward transmission.

The transfer of data through the Terachannel fabric is packaged in cells termed *tensors*. An arbitration cycle transfers one tensor per port through the StarCeptor. Each tensor consists of 6-8 *vectors*, each of which is transferred though the StarCeptor in one system clock cycle. A vector consists of one byte per plane in the StarCeptor.

The sizes of the vector and tensor for a particular application is determined by the bandwidth required in the fabric and the most appropriate cell size. The following sections show the typical packaging of the data as it flows through the Terachannel fabric for ATM and GbE:

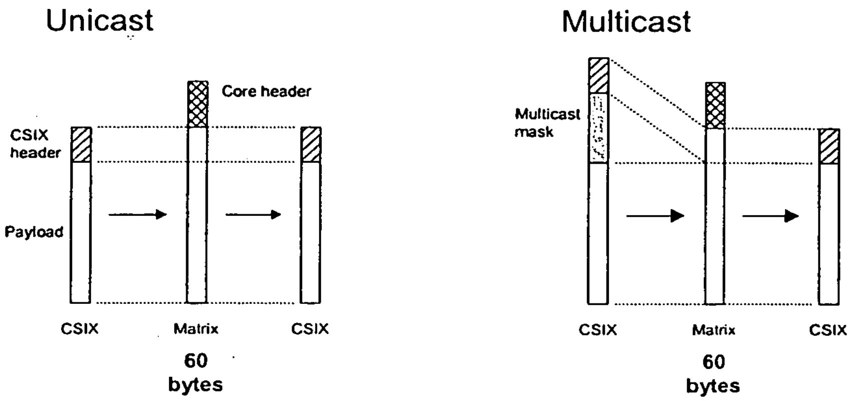
- **Fixed length ATM cells**

This demonstrates the packaging of 53 byte ATM cells into 60 byte tensors (6 vectors of 10 bytes).

The header and payload length are as follows:

	Size (bytes)	**Contents**
CSIX header	4	Frame_type (4), Channel (4), Priority (4), Destination_Address (12), Length (8)
Core header	3	Target_TM (8), TEC (16)
Multicast mask (optional)	16	Mask (128)
Payload	53	ATM cell

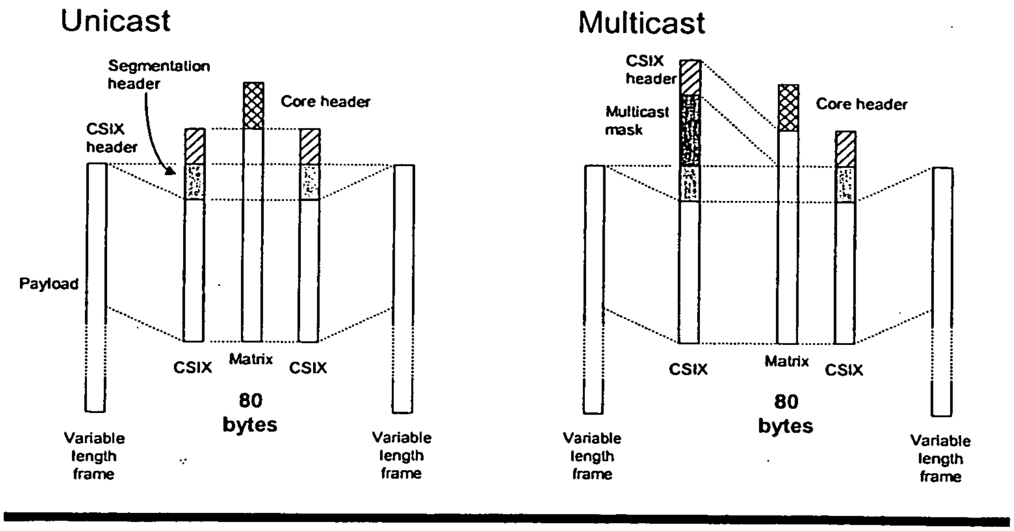
The ingress StarRouter analyses the CSIX header and wraps the CSIX packet with the Core header to create a 60 byte tensor in an ingress queue. When the StarMaster grants the required connection the tensor passes through the StarCeptor in one switch cycle to the egress StarRouter which writes the tensor into the egress queue indicated in the core header. When the tensor reaches the head of the egress queue, the core header is stripped off and the remaining CSIX packet is sent to the egress Traffic Manager.
If the CSIX frame type indicates a multicast packet, the ingress StarRouter strips out the multicast mask and replicates the packet into the indicated ingress queues, modifying the Target_TM field for each copy as appropriate. The flow then proceeds as for unicast, except that the tensor is written simultaneously into multiple egress buffers after passing through the StarCeptor.



- **Variable length packets**

In the case of a variable length packet switch (for example, for Gigabit Ethernet), a Traffic Manager with segmentation and reassembly, converts the variable length packets into CSIX packets at ingress, embedding the SAR header in the payload. The CSIX packets are then transported though the Terachannel fabric in the same way as for the ATM example above, except that the tensor size is set to 80 bytes (8 vectors of 10 bytes) allowing up to 70 bytes of ethernet frame to be carried in a single segmented packet:

	Size (bytes)	Contents
CSIX header	4	Frame_type (4), Channel (4), Priority (4), Destination_Address (12), Length (8)
Core header	3	Target_TM (8), TEC (16)
Multicast mask (optional)	8	Mask (64)
Segmentation header (eg.,)	3	Type, Counter, Address, Length
Payload	Up to 70	Ethernet frame (or part of)



Note that the segmentation header is private to the TCs and is shown for illustrative purposes only. The Terachannel treats it transparently as part of the payload.
The CSIX interface description allows for truncated packets, that is, if a TM has less payload than would fill a tensor it can send a shortened CSIX packet. The ingress StarRouter only needs to store the short packet in the ingress queues (on fixed tensor boundaries). The fixed size tensors will then have the invalid bytes discarded at the egress StarRouter.

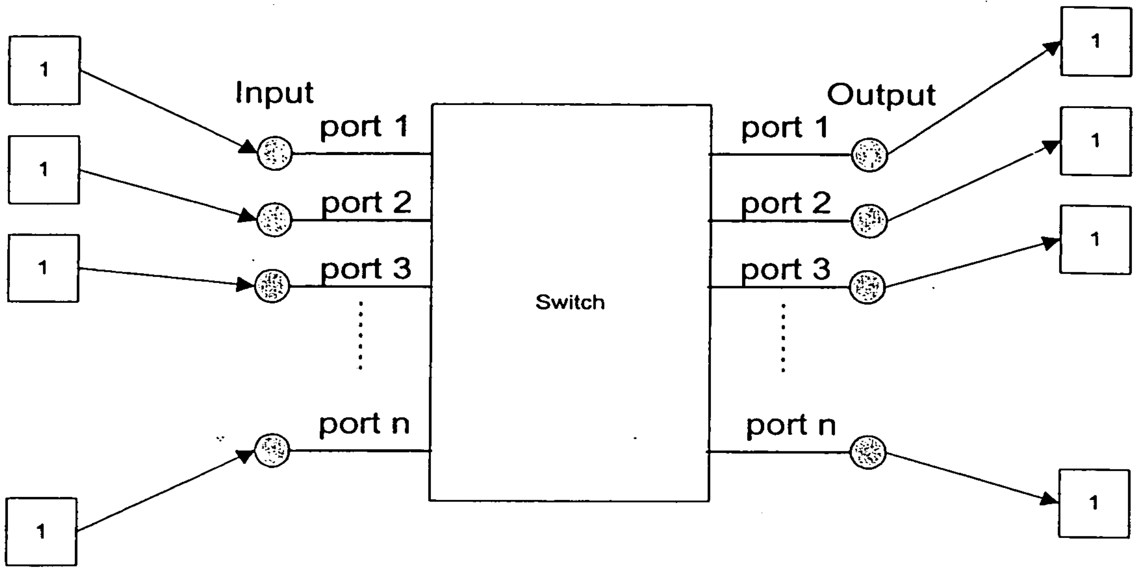


FIG. 1: Architecture

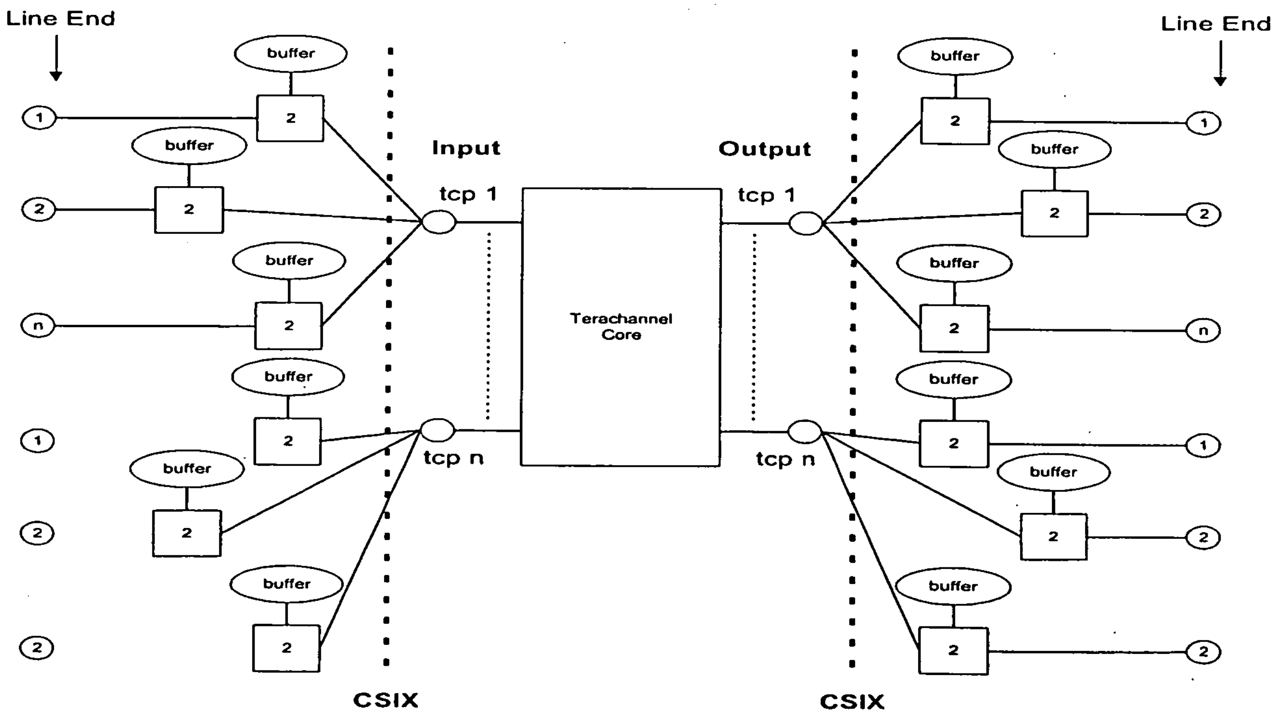


FIG. 2 : Switch Core

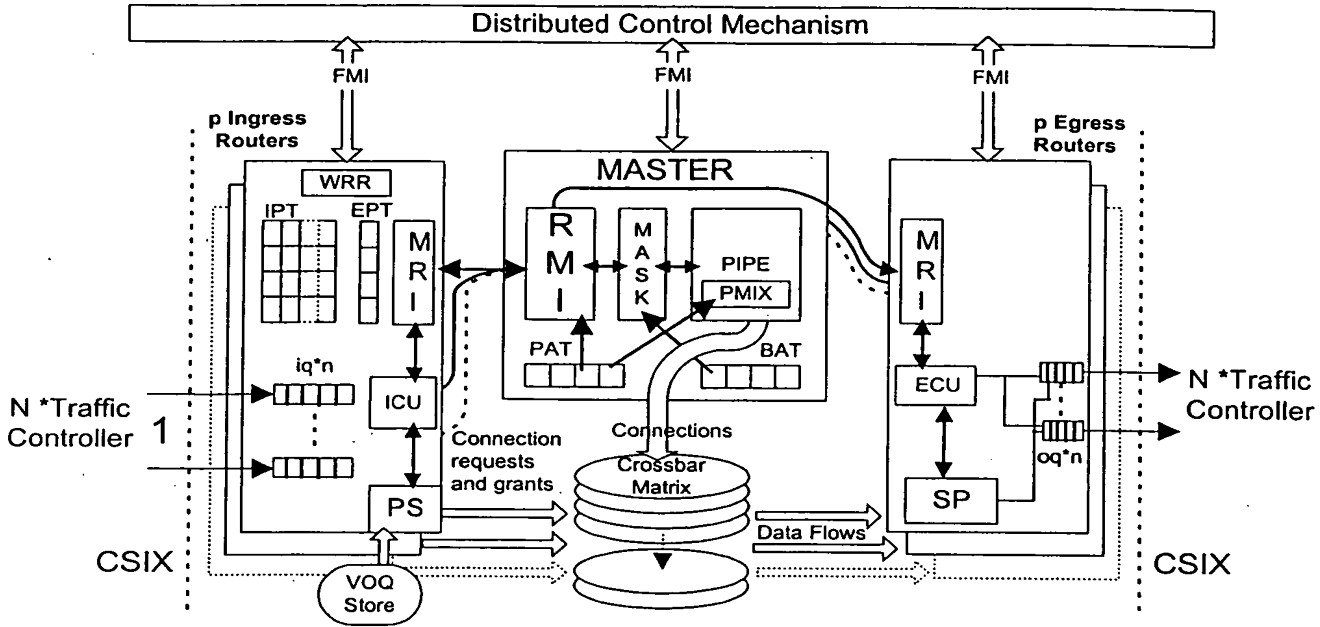


FIG 3: Teracore

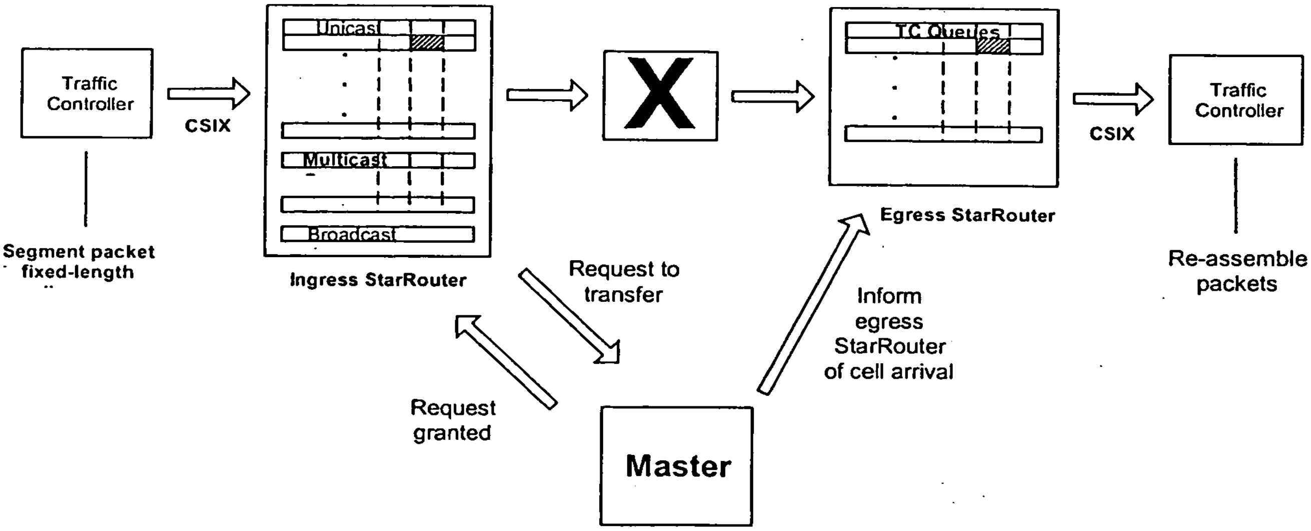


FIG 4:
Flow of data through the switch fabric